

# Dip Test Distributions, P-values, and other Explorations

Martin Mächler  
ETH Zurich

---

## Abstract

...  
...

*Keywords:* MPFR, Arbitrary Precision, Multiple Precision Floating-Point, R.

---

## 1. Introduction

FIXME: Need notation

$D_n := \text{dip}(\text{runif}(n))$ ;

but more generally,

$$D_n(F) := D(X_1, X_2, \dots, X_n), \quad \text{where } X_i \text{ i.i.d. }, X_i \sim F. \quad (1)$$

Hartigan and Hartigan (1985) in their “seminal” paper on the dip statistic  $D_n$  already proved that  $\sqrt{n} D_n$  converges in distribution, i.e.,

$$\lim_{n \rightarrow \infty} \sqrt{n} D_n \stackrel{\mathcal{D}}{=} D_\infty. \quad (2)$$

A considerable part of this paper is devoted to explore the distribution of  $D_\infty$ .

## 2. History of the diptest R package

Hartigan (1985) published an implementation in Fortran of a concrete algorithm, where the code was also made available on Statlib<sup>1</sup>

On July 28, 1994, Dario Ringach, then at NY University, asked on Snews (the mailing list for S and S-plus users) about distributions and was helped by me and then about `dyn.load` problems, again helped by me. Subsequently he provided me with S-plus code which interfaced to (a `f2ced` version of) Hartigan’s Fortran code, for computing the dip statistic. and ended the (then private) e-mail with

---

<sup>1</sup>Statlib is now a website, of course, <http://lib.stat.cmu.edu/>, but then was *the* preferred way for distributing algorithms for statistical computing, available years before the existence of the WWW, and entailing e-mail and (anonymous) FTP

I am not going to have time to set this up for submission to StatLib. If you want to do it, please go ahead.

Regards, Dario

- several important bug fixes; last one Oct./Nov. 2003

However, the Fortran code file <http://lib.stat.cmu.edu/apstat/217>, was last changed Thu 04 Aug 2005 03:43:28 PM CEST.

We have some results of the dip.dist of *before* the bug fix; notably the “dip of the dip” probabilities have changed considerably!!

- see rcs log of ../../src/dip.c

### 3. 21st Century Improvement of Hartigan<sup>2</sup>'s Table

((

Use listing package (or so to more or less “cut & paste” the nice code in ../../stuff/new-simul.Rout-1e6

))

### 4. The Dip in the Dip's Distribution

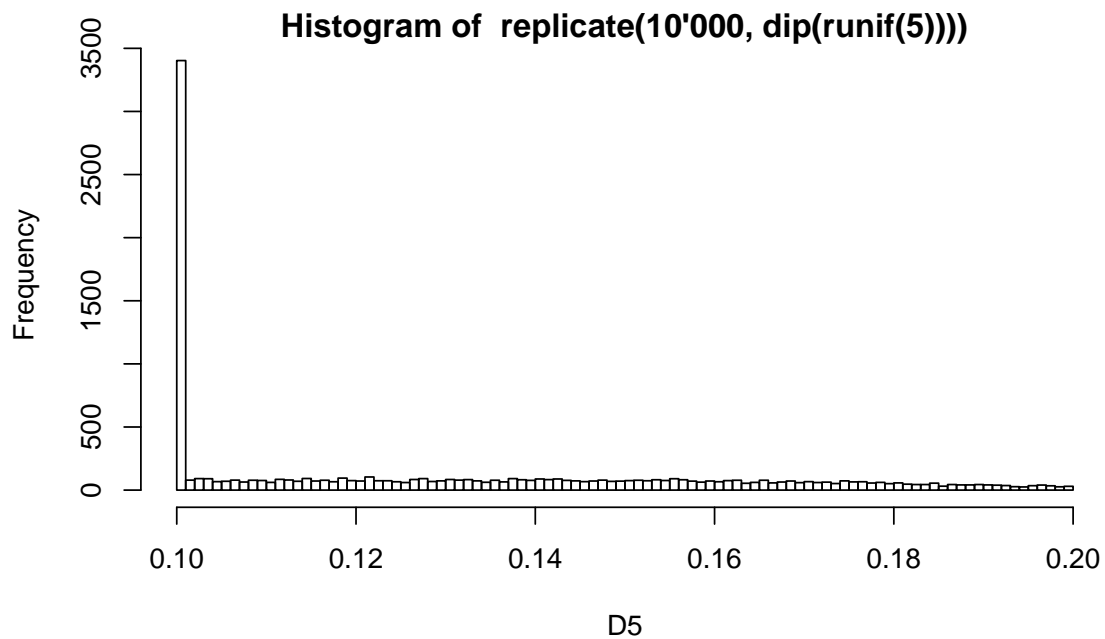
We have found empirically that the dip distribution itself starts with a “dip”. Specifically, the minimal possible value of  $D_n$  is  $\frac{1}{2n}$  and the probability of reaching that value,

$$P \left[ D_n = \frac{1}{2n} \right], \quad (3)$$

is large for small  $n$ .

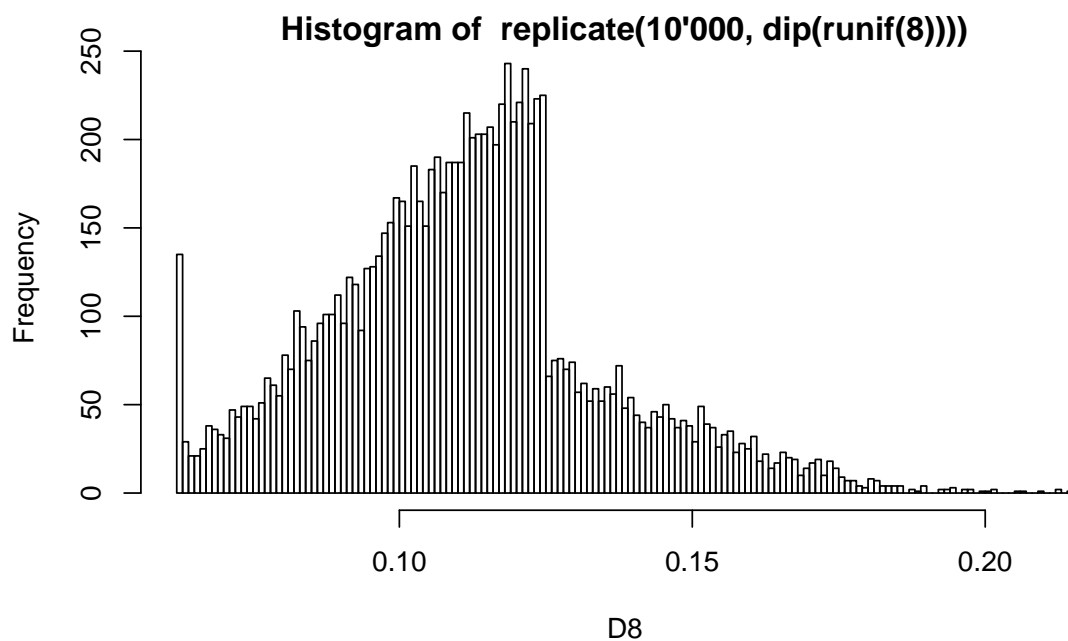
E.g., consider an approximation of the dip distribution for  $n = 5$ ,

```
R> require("diptest") # after installing it ..
R> D5 <- replicate(10000, dip(runif(5)))
R> hist(D5, breaks=128, main = "Histogram of replicate(10'000, dip(runif(5)))")
```



which looks as if there was a bug in the software — but that look is misleading! Note how the phenomenon is still visible for  $n = 8$ ,

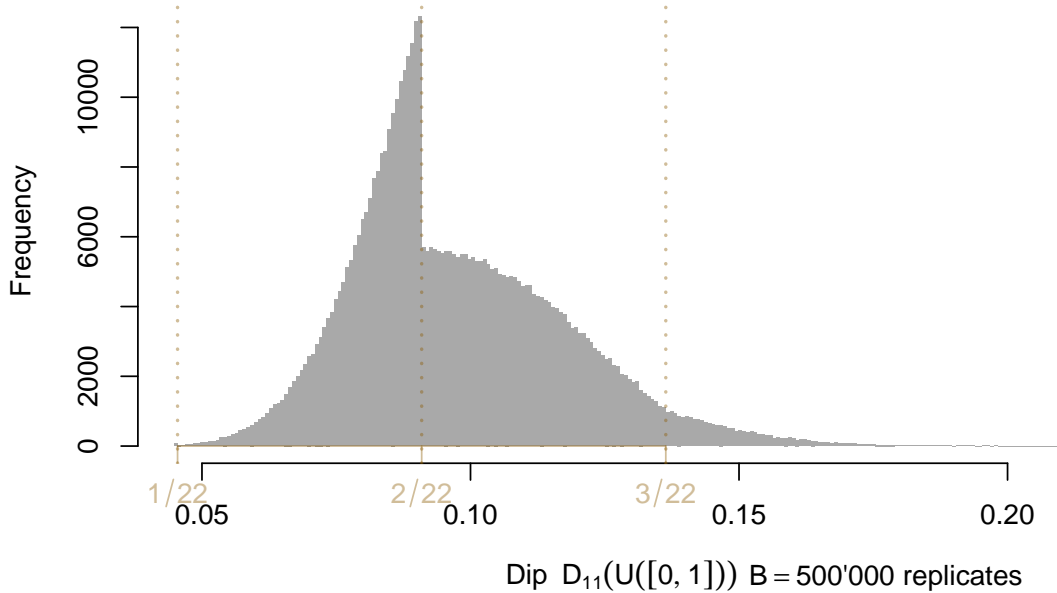
```
R> D8 <- replicate(10000, dip(runif(8)))
R> hist(D8, breaks=128, main = "Histogram of replicate(10'000, dip(runif(8))))")
```



Note that there is another phenomenon, in addition to the point mass at  $1/(2n)$ , particularly visible, if we use *many* replicates,

```
R> set.seed(11)
R> n <- 11
```

```
R> B.s11 <- 500000
R> D11 <- replicate(B.s11, dip(runif(n)))
```



FIXME:

use ‘`../../stuff/sim-minProb.R`’  
and ‘`../../stuff/minProb-anal.R`’

Further, it can be seen that the *maximal* dip statistic is  $\frac{1}{4} = 0.25$  and this upper bound can be reached simply (for even  $n$ ) using the the data  $(0, 0, \dots, 0, 1, 1, \dots, 1)$ , a bi-point mass with equal mass at both points.

## 5. P-values for the Dip Test

Note that it is not obvious how to compute  $p$ -values for “the dip test”, as that means determining the distribution of the test statistic, i.e.,  $D_n$  under the null hypothesis, but a natural null,  $H_o : F \in \{F_{\text{cadlag}} \mid f := \frac{d}{dx} F_{\text{isunimodal}}\}$  is too large. Hartigans’(1985) argued for using the uniform  $U[0, 1]$  i.e.,  $F'(x) = f(x) = \mathbf{1}_{[0,1]}(x) = [0 \leq x \leq 1]$  (Iverson bracket) instead, even though they showed that it is not quite the “least favorable” one. Following Hartigans’, we will define the  $p$ -value of an observed  $d_n$  as

$$P_{d_n} := \mathbb{P}[D_n \geq d_n] := \mathbb{P}[\text{dip}(U_1, \dots, U_n) \geq d_n], \quad \text{where } U_i \sim U[0, 1], \quad \text{i.i.d.} \quad (4)$$

### 5.1. Interpolating the Dip Table

Because of the asymptotic distribution,  $\lim_{n \rightarrow \infty} \sqrt{n} D_n \stackrel{\mathcal{D}}{=} D_\infty$ , it makes sense to consider the “ $\sqrt{n}D_n$ ”-scale, even for finite  $n$  values:

```
R> data(qDiptab)
R> dnqd <- dimnames(qDiptab)
R> (nn. <- as.integer(dnqd[["n"]]))
```



```
R> y.0 <- sqrt(n0)* qDiptab[i.n ,]
R> y.1 <- sqrt(n1)* qDiptab[i.n+1,]
R> (Pval <- 1 - approx(y.0 + f.n*(y.1 - y.0),
                     P.p,
                     xout = sqrt(n) * D)[["y"]])
```

```
[1] 0.001809527
```

```
R> ## 0.018095
```

Finally, in May 2011, after several years of people asking for it, I have implemented a `dip.test` function which makes use of a — somewhat more sophisticated — interpolation scheme like the one above, to compute a  $p$ -value. As `qDiptab` has been based on  $10^6$  samples, the interpolation yields accurate  $p$ -values, unless in very extreme cases. Here is the small ( $n = 63$ ) example from Hartigan<sup>2</sup>,

```
R> data(statfaculty)
R> dip.test(statfaculty)
```

```
Hartigans' dip test for unimodality / multimodality
```

```
data: statfaculty
D = 0.059524, p-value = 0.08672
alternative hypothesis: non-unimodal, i.e., at least bimodal
```

where, from a  $p$ -value of 8.7%, we'd conclude that there is not enough evidence against unimodality.

## 5.2. Asymptotic Dip Distribution

We have conducted extensive simulations in order to explore the limit distribution of  $D_\infty$ , i.e., the limit of  $\sqrt{n} D_n$ , (2).

Our current R code is in ‘ `../../stuff/asymp-distrib.R` ’ but the simulation results (7 Megabytes for each  $n$ ) cannot be assumed to be part of the package, nor maybe even to be simply accessible via the internet.

## 6. Less Conservative Dip Testing

## 7. Session Info

```
R> toLatex(sessionInfo())
```

- R Under development (unstable) (2016-12-05 r71733), x86\_64-pc-linux-gnu
- Locale: LC\_CTYPE=en\_US.UTF-8, LC\_NUMERIC=C, LC\_TIME=en\_US.UTF-8, LC\_COLLATE=C, LC\_MONETARY=en\_US.UTF-8, LC\_MESSAGES=C, LC\_PAPER=en\_US.UTF-8, LC\_NAME=C, LC\_ADDRESS=C, LC\_TELEPHONE=C, LC\_MEASUREMENT=en\_US.UTF-8, LC\_IDENTIFICATION=C
- Running under: Debian GNU/Linux stretch/sid

- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: diptest 0.75-7
- Loaded via a namespace (and not attached): compiler 3.4.0, tools 3.4.0

## References

Hartigan JA, Hartigan PM (1985). “The Dip Test of Unimodality.” *Annals of Statistics*, **13**, 70–84.

Hartigan PM (1985). “Computation of the Dip Statistic to Test for Unimodality.” *Applied Statistics*, **34**, 320–325.

### Affiliation:

Martin Mächler

Seminar für Statistik, HG G 16

ETH Zurich

8092 Zurich, Switzerland

E-mail: [maechler@stat.math.ethz.ch](mailto:maechler@stat.math.ethz.ch)

URL: <http://stat.ethz.ch/people/maechler>